Shared meaning in children's evaluations of art: A computational analysis

Astrid Schepman, Julie Kirkham, Paul Rodway, Jordana Lambert and Anastasia Locke

University of Chester

Author Note

Department of Psychology, University of Chester.

The data acquisition work on which this analysis is based was funded by an Internal Grant by the University of Chester to the first three authors. Brian Rodway (brian@affinitystudios.co.uk) of Affinity Studios, UK, (http://www.affinitystudios.co.uk/index.html), wrote the Javascript software that called the semantic similarity service and stored the scores.

Correspondence concerning this article should be addressed to Dr Astrid Schepman, Department of Psychology, University of Chester, Parkgate Road, Chester, Cheshire, CH1 4BJ, United Kingdom. Email: a.schepman@chester.ac.uk

Abstract

Art appreciation is often considered highly individual, but research has shown that there is also a shared element, which may be due to shared meanings and associations triggered by artworks. In the current analysis, we examined semantically based justifications given to aesthetic evaluations of abstract and representational artworks provided by 80 primary schoolchildren, aged 4, 5, 8, and 10 years. Using a computational semantic similarity analysis technique (UMBC Ebiquity), the authors found that children showed evidence for shared meaning in response to representational but not abstract art. The effect was present from age 4 through to age 10. In addition, it was found that the presence of semantic elements in the justifications boosted aesthetic appreciation, especially of abstract artworks. This suggests that individually constructed meaning is key to aesthetic appreciation and is, to an extent, independent from the meaning that might be assumed to be inherent in artworks, particularly if it is representational. The authors evaluate their findings in relation to aesthetic and developmental theories and make suggestions for future research. They argue that the current data, alongside calibrating analyses that apply their randomization and semantic analysis protocol to children's picture naming responses, further demonstrate the robustness of the computational semantic similarity analysis method, with great potential for further studies in semantic interpretation of art or other types of stimuli.

Keywords: Empirical aesthetics; Shared understanding; Art appreciation; Child Development; Computational Semantic Analysis

Shared meaning in children's evaluations of art: A computational analysis

The aesthetic appreciation of art is commonly thought to be a subjective matter (McManus, 1980; Leder, Gerger, Dresser & Schabmann, 2012), driven by the formal artistic properties of the artwork and a variety of individual factors, including the viewer's cognitive, emotional and associative reactions to the artwork (e.g. Augustin, Leder, Hutzler, & Carbon, 2008; Locher, Krupinski, Mello-Thoms, & Nodine, 2007; Pelowski, Markey, Lauring, & Leder, 2016). However, research has also demonstrated that there can be a high level of similarity in taste for artworks across individuals (Eysenck, 1940; Vessel & Rubin, 2010) in addition to a degree of individuality. This raises an important issue in empirical aesthetics (Palmer, Schloss & Sammartino, 2013; Leder, Goller, Rigotti, & Forster, 2016; Vessel & Rubin, 2010), which concerns the extent to which shared taste is mediated by internal cognitive processes arising from shared experiences with the particular attributes of the artwork (Leder, et al. 2016; Vessel & Rubin, 2010).

Researchers have begun to establish which factors lead to shared taste, and have identified that shared taste is greater for representational art than for abstract art (Leder, Goller, Rigotti & Forster, 2016; Schepman, Rodway, Pullen & Kirkham, 2015b; Vessel & Rubin, 2010). Leder et al. (2016) also showed that shared taste is lower for abstract art than for facial attractiveness and that people have a weak understanding of what others find attractive in abstract art (Leder et al. 2016). A reasonable inference that can be made from these findings is that it is the meaning of the subject matter depicted in representational artwork that is the root cause of the increase in shared liking (Vessel & Rubin, 2010). For example, an artwork depicting a holiday scene with sunbathers and ice cream may elicit a higher level of shared (positive) taste in a set of viewers than an abstract artwork using similar colors and tones but without recognizable content, because the former elicits similar

shared (pleasant) thoughts, while the latter is more likely to elicit fewer shared thoughts (see Leder et al., 2016). However, this assumption cannot necessarily be inferred from the presence of representational meaning in the artwork without specific evidence that the meaning engendered by viewing artworks is shared across individuals. Schepman, Rodway and Pullen (2015a) were the first researchers to use a computational semantic analysis technique to examine the shared meaning for representational and abstract artworks directly. Schepman et al. (2015b) had collected participants' associative responses to artworks and had asked viewers to rate these responses for valence ((very) positive, neutral, (very) negative). They had found that these valence ratings converged more across individuals in response to representational than abstract artworks. Schepman, Rodway and Pullen (2015a) randomly paired the responses with other responses to the same artwork for the purpose of computing similarity scores. These pairs were then put through semantic analysis software and it was found that the semantic similarity for the randomly paired responses was greater for representational than for abstract art. This showed, for the first time, greater shared meaning in representational than abstract art. It also provided firm empirical evidence supporting the notion that shared liking is due to shared meaning, rather than this simply being a reasonable inference.

The current work examined a similar question to that investigated by Schepman et al. (2015a), but from a developmental perspective. Adopting a developmental perspective was important because it served as a further test of the idea that shared meaning across individuals underpinned shared aesthetic appreciation. If shared meaning develops from shared experiences, and drives shared liking in aesthetics, then those shared experiences (and meanings) must take time to accumulate and develop. It might be expected that at younger ages children should have fewer shared experiences, less shared meaning, and lower levels of

shared taste for representational art, but shared taste and shared meaning should increase with age. Conversely it is probable that abstract art which lacks semantic content does not elicit shared experiences in viewers (Leder et al. 2016), and so shared taste for abstract art will not show a similar increase with age (Rodway et al. 2016).

Although a number of studies have examined the development of children's aesthetic appreciation, relatively little attention has been paid to researching children's aesthetic understanding (Freeman & Parsons; 2001; Schabmann, Gerger, Schmidt, Wogerer, Osipov & Leder, 2015). Studies that have been conducted have not always considered the influence of different styles of artwork or have used methodologies where children's understanding is inferred from basic task completion but is not explicitly articulated (e.g., in matching tasks; Carothers & Gardner, 1979; Blank, Massey, Gardner & Winner, 1984). Early descriptive studies in this area (e.g., Machotka, 1966; Parsons 1987) used interview data and basic frequency analysis to produce stage models of aesthetic development. These suggested an age-related progression from evaluations based on color and subject matter towards consideration of more sophisticated properties including style. Follow-up empirical research conducted by Lin and Thomas (2002) supported a general trend towards increasingly complex aesthetic evaluations from the age of four years to adulthood. However, this was dependent on individual factors such as the level of personal interest in art, art experience and the aspect of the art that was considered (e.g., color, subject matter, associations). These aspects were derived and coded from participants' justifications for a particular artwork that they had themselves selected from a range of five styles (including abstract, fine, contemporary, cartoon and humorous). Surprisingly, the pattern of responses across the different aspects was generally similar across all art styles (despite variation in representational content), although references to subject matter were lower for abstract art.

Although subject matter and associations were considered in Lin and Thomas's (2002) study, 'meaning or 'understanding' was not a category that emerged from participant's justifications. As Rodway, Kirkham, Schepman, Lambert and Locke (2016) note, this could be because participants' self-selecting of artworks reduced the range and complexity of justifications that were given. Alongside other variables including arousal and emotion, more recent research by Leder, Gerger, Dressler and Schabmann (2012) considered whether 16 to 62 year olds' liking and comprehension of art was influenced by decreasing levels of representational content shown sequentially in classical, modern and abstract artworks. Although not a developmental study, the authors divided their sample into 'low' and 'high' art expertise groups. Expertise is one factor which may increase in line with age, schooling and accumulation of life experiences (Leder et al., 2012). Overall, comprehension and liking (measured on a 1 to 9 point scale) was higher for the more representational artworks (classical and modern) than the abstract artworks, and this effect increased significantly with level of expertise.

Schabmann et al., (2015) extended this study into a developmental context by utilizing the same procedure with a sample of kindergarten (42 4-7 year olds) and school age children (52 9-11 year olds). Again, both liking and comprehension were significantly higher for both age groups for the representational (classical and modern) than the abstract artworks. Oral and written justifications for aesthetic preference (i.e. 'why did you find the artwork beautiful?') were collected for kindergarteners and school age children respectively. Categorization of the justifications and frequency analysis showed that the majority of children in both age groups referred to color and content to formulate their decisions, however the semantic content of these justifications was not explored, and no further inferential statistics were conducted.

Similarly to the studies by Leder et al., (2012) and Schabmann et al., (2015), Rodway et al., (2016) considered children's preferences for representational and abstract art using a quantitative rating scheme where participants were asked to indicate their liking of ten pictures of each art type by giving a rating from 1 to 5 stars (5 highest). Children were also asked to verbally explain the reasons for their preferences, and their responses were audio recorded, transcribed and then coded into categories. British children were sampled throughout the course of primary school age education at ages 4, 6, 8 and 10 years with twenty participants in each age group. This study found greater shared liking for representational art compared to abstract art in eight- and ten-year-old children, but no differences in liking between art type were found in four or six year olds. One possible interpretation for this finding is that children's aesthetic preferences mirror their own art production abilities (Tinio, 2013), with the representational content of children's pictures increasing with age and becoming more realistic by middle childhood (Jolley, 2010). However, even very young children may intend for their pictures to be representational, it is the lack of cognitive and motor skills that may prevent this intention from being realized (Toomela, 2002). An alternative interpretation, offered by Rodway et al., (2016), is that the divergence in shared liking for representational compared to abstract art from age eight upwards is the product of an accumulation of shared experiences, meanings, and associations in response to the depicted subject matter as childhood progresses. It could be these shared experiences in response to representational art that underpin shared liking, with no similar process available for abstract art, which may thus lead to more idiosyncratic evaluations. This explanation is supported by research showing the role of prior life experiences and their associations in the development of children's categorization abilities, specifically the movement from decisions based on basic level perceptual properties towards more complex

semantic judgements (Blaye, Bernard-Peyron, & Bonthoux, 2000; Murphy, 2002) which is further dependent on biological and cognitive maturation (Colunga & Smith, 2005; Qin, Cho, Chen, Rosenberg-Lee, Geary, & Menon, 2014).

Testing the hypothesis that shared liking is underpinned by shared meaning is the primary aim of the current study. Whilst other research (Schabmann et al., 2015; Rodway et al., 2016) has studied developmental trends in children's aesthetic preferences, semantic meaning has not been the focus of these studies. In the current article we analyzed a previously unanalyzed aspect of the data from Rodway et al.'s (2016) study, namely the semantically-based content of the verbal justifications that children had given for their ratings. Our aim was to establish whether shared semantic content was greater for representational than for abstract art in this sample of children, and, if so, from what age. Finally, we also examined whether the presence of meaning-based justifications led to more favorable evaluations.

Operationalization and Hypotheses

Conceptually, our general hypothesis was that there would be evidence of greater shared meaning for representational art than for abstract art. This was based on the findings of Schepman et al., (2015a) which used the same procedure but with an adult sample, and on the developmental findings of Schabmann et al., (2015), and on various studies which suggest that level of art expertise can influence style-related preferences (e.g., Gardiner, Winner & Kirchner, 1975; Augustin & Leder, 2006, Leder et al., 2012). In order to phrase the hypotheses more precisely in terms of dependent and independent variables, a brief outline of the operationalization of these variables is given first.

The dependent variable in this set of analyses was a semantic similarity score (Han, Kashyap, Finin, Mayfield, & Weese, 2013) ranging from 0 to 1, with higher scores indicating greater similarity. A detailed discussion will follow, but in brief for now, this dependent variable captured the extent to which meaning in the verbal responses to the artworks was shared. Specifically, a semantic similarity score captures the similarity between two tokens of text (in our case, two justifications for the liking of specific artworks). As a general example, the pair of statements "He likes the sea" and "She loves the ocean" would, by most people, be considered to overlap in meaning. The analysis method we use places a number on this, namely 0.633. In contrast, the statements "His popularity waned over the years" and "It is because there are five" show little, if any, semantic overlap, and the analysis method we use gives this a semantic similarity score of 0. These are not responses from our data, but simply illustrative examples created by us. As is also demonstrated by the examples, the semantic analysis software calculates a score for pairs of sentences. Thus, an important part of our analysis protocol involves pairing the children's responses for the purpose of applying the semantic similarity analysis, to enable inferential statistical analysis. This will be explained in more detail in the method, but in outline, we created sets of pairings of responses. To test our hypotheses, we had to create two different types of pairings. The first was of responses that different children had supplied for the same artwork. We called these pairs "experimental". We also created pairings of responses across the whole set of artworks within a particular type, and called these pairs "baseline control", because these controlled for the possibility that there may be baseline differences in linguistic aspects of the responses as a function of art type, which could form a confound. The type of pairing (experimental vs. baseline control) formed the key independent variable in our analysis.

Having explained the key variables, we return to our hypotheses, which can now be phrased more specifically in terms of these variables. Our first more precisely phrased hypothesis was that, if meaning is shared for representational, but not for abstract artworks, the similarity scores of experimental justification pairings should exceed the semantic similarity scores for baseline pairings for representational artwork, but this should not apply to abstract artworks.

A second hypothesis, which tested our interpretation of Rodway et al.'s (2016) finding that shared liking started emerging at age eight for representational art only used experimental and baseline control pairings as described above, but separated by age. For our interpretation that the onset of shared liking is due to an onset of shared meaning at age eight to be supported, we would expect the experimental pairings to exceed the baseline control pairings for representational artworks at ages eight and ten, but not at ages four and six. Note that, alongside testing the specific hypothesis, this additional set of pairings also provided an opportunity to replicate the findings for our first hypothesis and Schepman et al. (2015a), as explained in more detail later.

Further, we also carried forward the numerical aesthetic evaluations by the children as part of Rodway et al.'s (2016) study. In Rodway et al. (2016) we had not found any differences in aesthetic evaluations as a function of art type (abstract vs. representational). In the current article we explored whether there would be differences in aesthetic evaluations as a function of the presence or absence of semantically-based justifications provided by the children. We expected artworks attracting comments related to meaning to be rated more favorably than those with non-semantic justifications only, in line with proposals by, among others, Landau, Greenberg, Solomon, Pyszczynski, and Martens, (2006) and Russell (2003)

that meaning heightens aesthetic-hedonic value. Note that in this analysis meaning is defined as subjective, rather than as inherent in the stimulus.

Method

Data Acquisition and Coding

The process of data acquisition is described in Rodway et al. (2016), but is repeated here for ease of reference, in slightly edited form to supply some additional details. The participants were primary school children attending an English national curriculum school. Based on national statistics, this school had average achievement levels, a somewhat lower than national average proportion of non-native speakers (indicating relatively low ethnic diversity) and approximately half the national average proportion of children eligible for free school meals (indicating relatively low poverty). The children were recruited via the school, following parental opt-out consent and child assent, with the teachers facilitating the selection process, selecting children who had no specific interest in or experience with visual arts. There were eighty children in total, 20 in each age group (aged 4, 6, 8 and 10). Characteristics of the sample were as follows: 4-year-olds (mean age 4.7, SD 3 months; 15 males and 5 females); 6-year-olds (mean age 6.4, SD 4 months; 8 males, 12 females); 8-year-olds (mean age 8.7, SD 3 months; 10 males, 10 females); 10-year-olds (mean age 10.6; SD 4 months; 10 males, 10 females). Twenty artworks (ten representational, ten abstract, see Supplementary Data for full list and URLs) were presented in color-printed form, on white A4 paper, with one image per page. Booklets were created which displayed the set of images in one of three random orders. All participants were tested individually by author JK at a desk

in a quiet room next to their usual classroom. The study was introduced to each child as follows: "*We see pictures every day, for example in books and on walls at school and at home. We may like some pictures more than we like others. Today I am interested in what pictures you like. There is no right or wrong answer; I just want you to tell me what you think of each picture that I show you*." Participants were also instructed that they could ask for a break at any time during the procedure. Each artwork in the booklet was presented sequentially to the participants with the following instruction repeated for each of the 20 artworks *"I would like to know how much you like this picture. Would you give it 1 star (you don't like the picture at all), 2 stars (you think the picture is ok but that are some parts that you don't like), 3 stars (the picture is good. You like it), 4 stars (the picture is very good. You like it a lot) or 5 stars? (The picture is excellent. You love it)."* Participants were then instructed to point to the number of stars that they wanted to give the artwork on the star rating sheet. If necessary the instructions were repeated. After indicating their rating for each individual artwork participants were then asked "why did you give that picture [1–5] stars?" The instruction was repeated as necessary to elicit a response and any queries raised by the participants were answered as follows: *"I am interested to know the reasons why you gave this picture 1–5 stars. There is no right or wrong answer. I only want to know what you think about the picture."* Due to the potentially limited verbal abilities of some of the children, three additional categories of prompts were used by JK to support and clarify participant's responses. Firstly, for basic responses without any explanation (for e.g., "I like it") participants were prompted by asking "why?" or "what?" questions to elicit further detail. Secondly, if participants were explaining a concept but were unable to retrieve the appropriate word to describe it (or used the incorrect word), then JK provided the correct word (e.g., "calf"). Finally, if participants provided an explanation with reference to part of a

picture but it was unclear what part they were referring to, JK asked participants to clarify this. To keep the time frame of the study manageable for primary school children each of the 20 artworks was presented to the participant for a maximum of five minutes.

Subsequently, authors JL and AL transcribed and coded the justifications, using a coding scheme with initial codes from the literature (e.g., Machotka, 1966; Parsons, 1987; Lin and Thomas, 2002) as well as from JK's direct experience of listening to the participants' responses. Following practice coding by authors JL and AL, these were reduced to 14 final categories. Of these, four made reference to the content or meaning of the artwork, namely *Subject Matter* (any reference to the content or objects perceived to be depicted in the artwork), *Associations* (reference to entities of which the rater was reminded by the artwork), *History / Culture* (reference to historical or cultural entities to which the artwork was related by the rater) and *Emotion / Mood* (to catch justifications that referred to emotions conveyed by the painting, in the event mostly related to the subject matter – e.g. "She looks happy".) The ten remaining categories did not refer to the content or meaning of the artwork, but to other aspects. Two related to visual aspects of the artwork, namely *Formal Artistic Properties* (referring to e.g. lines, composition, style) and *Color* (any reference to color). There was a category to capture reference to the artwork's production, namely *Artist* (reference to the artist as an agent), and there were two for the process of understanding the artwork, namely *Understanding / Interpretation* (the process of understanding - or not understanding -, but not the content of the work itself), and *Perceptual Fluency* (the ease or difficulty of interpreting the artwork), as well as general categories, namely *Interest* (whether or not the artwork was interesting); *Comparison* (how the artwork compared to other artworks in terms of rating), *Function* (uses for the artwork, e.g. hanging on a wall), *Basic Liking* (simple expressions of liking or dislike for the artwork without other elements of

justification), *Other* (a category used only if no other categories applied). The two independent coders categorized all 1600 justifications into one or multiple categories that applied to the response. Codes were compared and any simple data entry errors were repaired. The remaining codes which showed initial disagreements were resolved by discussion. Analysis of all 1600 justifications showed that in 92.9% of the justifications, the raters agreed on all 14 codes chosen, while in 7.1% of justifications, they needed to discuss one or more codes to reach agreement on the overall coding of the justification. This shows a high level of reliability. The codes where there was initial disagreement were settled by discussion between the two coders.

Data Selection

For the current set of analyses, we separated the justifications that had been coded as based at least in part on the meaning attributed by the child to the artwork from those not based in meaning. We used clear criteria by which to perform this separation. Authors AS, JK and PR carefully inspected all codes given by the coders and identified by consensus that responses coded with at least one of the codes *Subject Matter, Associations, Mood / Emotion, History / Culture* contained references to the meaning, content or subject matter of the artwork. Thus, responses that attracted at least one of these codes were included in our semantically-based set. Responses that had not been classified as containing elements compatible with these four codes, were deemed to reflect reactions that were not based on the meaning of the artworks, but on other aspects, such as the visual appearance of the artwork, the production process, the viewer's observation and interpretation process, the artwork's utility, or simply very basic and vague reasons for liking or disliking the artworks, in line

with the coding categories described above and in Rodway at al. (2016). It was felt that making the semantic vs. non-semantic division based on the prior coding categories was a strength, because the coding had been subject to careful validation and reliability checking.

Data Screening in Preparation for Semantic Analysis

Basic statistical inventory information was computed to check whether our intended analyses were viable, and for this there had to be enough responses per age group and art type that met the inclusion criteria. Table 1 shows that this was the case. We note that there were approximately 3.4 times as many semantic responses to representational artworks than to abstract artworks. We evaluate any impacts of this imbalance in numbers on the statistical analysis in the Results section using a control analysis. Before we discuss the processing of these data further, we discuss key aspects of the method in more detail.

--- INSERT TABLE 1 ABOUT HERE ---

Semantic analysis software and creating random pairs of responses: Principles and validity test

In order to analyze the semantic similarity between responses, we made use of pre-existing software that is able to quantify the semantic similarity between two sentences, words or phrases (http://swoogle.umbc.edu/SimService/index.html, see Han et al., 2013 for a full description of the algorithms, databases and validity / performance testing used during its

development). This software was also used in Schepman et al. (2015a). The UMBC eBiquity software combines distributional similarity, latent semantic analysis (see e.g. Landauer, Foltz, & Laham, 1998), and thesaurus methods (Wordnet; Miller, 1995) to calculate semantic similarity scores for pairs of words, phrases, or sentences, using multi-layered routines to optimize the accuracy of the semantic similarity scores. As described in Han et al. (2013), it performs well against other similar software, being the top software of its type in an annual competition that year. The software has three variants, of which we chose Semantic Textual Similarity (http://swoogle.umbc.edu/StsService/index.html) because it was better able than the other two software variants to handle longer responses of the type we observed in the justifications given by the children. Our choice of software was in part also informed by its availability for use and automatic querying by an Application Programming Interface, to make the processing of bulk data practical. In addition, using software that had been used in earlier work (Schepman et al. 2015a) helped us make direct comparisons between our two studies.

Semantic Textual Similarity (STS) software is under continuous development, and it is possible that additional software will be available in the future (see e.g. Vo & Popescu, 2016, for further recent developments). However, following major design improvements, their software "consistently and stably performs at the state of the art or top-tier level on all STS datasets from 2012 to 2015" (Vo and Popescu, 2016, p. 64), indicating that software developed during 2012-2015 already showed high levels of performance upon which it was difficult to improve substantially. Han et al. 2013 was the the top-tier system for 2013, and therefore our current use of the Han et al. (2013) software is likely to lead to optimal results, given the state of the art in STS software development.

The child participants gave single responses to each artwork, so we had to include a stage of data processing in which we created pairs of responses to be able to use the semantic analysis software (see e.g. Koch, Alves, Krüger, & Unkelbach, 2016; Potter, Corneille, Ruys, & Rhodes, 2007; Unkelbach, Fiedler, Bayer., Stegmüller, & Danner, 2008 for the use of pairings to acquire similarity data). We achieved this by creating contrasting sets of random pairings, as briefly mentioned in the Introduction. In one randomization, the "experimental" randomization, responses to an artwork were paired with a randomly selected response by another participant to the same artwork, while in a different randomization, the "baseline control" randomization, a response was paired with a randomly selected response to any artwork from the entire set. For example, if children viewed an artwork depicting a dog, it was likely that other children's responses to this artwork were more similar than other children's responses to an artwork depicting a woman laughing. This basic principle underlies our two sets of randomizations. A statistically significantly higher semantic similarity score for the experimental (within artwork) pairings than the baseline control (across artwork) pairings would be evidence of shared meaning. Note that the baseline is necessary as a comparison. The specific choices of baseline control conditions, using a different baseline for each art type, controlled for potential baseline linguistic differences between the responses to the different types of artwork, which could have formed a confound.

When the semantic similarity software is presented with a pair of entries, it computes a number between 0 and 1, of which 0 means no semantic similarity, or an item is not available in the database, while 1 indicates the highest level of semantic similarity. To give some examples from our set of response pairs, two randomly paired entries "Cos I like the doggy" paired with "I like it cos the dog's fetching the thing" had a similarity score of 0.814, while the pair "It is quite funny" with "Cos somebody's laughing" produced a similarity score

of 0.570, and the pair "The houses are too far away" with "Because it looks really cold" gave rise to a similarity score of 0.100. To compute semantic similarity scores for the entire set of random pairings, the pairs were sent through the semantic similarity software using a custom-made Java script (see Acknowledgements), which placed the results into a spreadsheet, ready for further analysis.

To validate our method, and with the further aim to explore a set of scores against which to compare our results, we undertook a calibrating analysis, which we report first. In this, we wanted to establish what the semantic similarity scores would be if recognizable objects were named, to establish measures of central tendency and dispersion as well as an inferential statistical difference resulting from the application of our randomization and semantic similarity analysis protocols to picture naming. We found suitable published picture-naming responses from 5-7 year-old children (Cycowicz, Friedman, Rothstein, & Snodgrass, 1997). Cycowicz et al. report children's naming responses to line drawings of 400 objects from three different previously published image sets. Of these three image sets, Snodgrass and Vanderwart's (1980) 260-item set was the largest. It was reported to also be the most reliable in terms of object recognition and name familiarity and for this reason we selected items only from this database. To determine a suitable sample size to select from the picture-naming database, we tried to match as closely as possible the total number of items in our own dataset of children's art meaning-based responses. In ours, there were 763 responses in total that were entered into the semantic analysis, as set out shortly. To put this analysis on a comparable footing in terms of number of trials, we randomly selected 28 items, each of which had 30 responses, giving 840 potential trials, of which we knew a proportion was missing due to children not being able to name the object. We used www.random.org to randomly select the following items, listed here preceded by their item numbers from the

original database: *4 anchor, 28 bird, 32 bottle, 48 carrot, 52 chain, 72 desk, 74 doll, 75 donkey, 81 duck, 88 finger, 91 flower, 97 fork, 114 hammer, 116 hanger, 126 kangaroo, 129 kite, 138 light bulb, 155 nose, 165 peanut, 182 rabbit, 186 rhinoceros, 190 rolling pin, 203 shirt, 221 suitcase, 235 toe, 243 trumpet, 247 vest, 251 watering can*. (Note that we replaced four items from an initial selection because there appeared to be more than 30 responses for those items, which may have indicated double coding of a type that was not immediately transparent to us). For our selection of items, children showed 73% name agreement, while not knowing the name and not knowing the object came to 4% each (thus accounting for 8% missing data), and providing alternative names accounted for 19% of the responses (e.g. "bunny" for "rabbit"), yielding a total of 772 responses, close to our own justification dataset of 763 items. Images and details of the responses given can be found in Cycowicz et al., (1997). For each item, the list was fully expanded so that each valid response given by each child was listed in a spreadsheet, with missing responses being omitted. Then, using www.random.org's random sequence generator, the responses were randomly paired in two ways: 1) experimental (within items): each naming response was randomly paired with a naming response from the set of responses to that same picture item; and 2) baseline control (across items): each verbal response was randomly paired with a response from the entire dataset, pooling responses across all items. This was done as one cycle only, not as an iterative or bootstrapping process. This followed the same randomization principle as for the responses to the artworks, to be reported shortly.

The random pairs of responses were presented to Han et al.'s (2013) semantic similarity software, and each pair was assigned a value ranging from 0 (no match or item not in the database) to 1 (maximum match). The scores for the entire two sets of random pairs were entered into statistical analyses with randomization (experimental vs. baseline control)

as the independent variable. The mean semantic similarity score for the experimental within-items pairings was .87 (SD = .24, SE = .01, 95% CI [.86, .89]), while for the control across-items pairings, this was .08 (SD = .19, SE = .01, 95% CI [.07, .09]). It is not surprising that that difference was significant on a Mann-Whitney test, chosen due to non-normality of the distributions, $Z = -33.63$, $p < .001$. What this analysis tells us is that if children of this age group are naming relatively recognizable drawings of objects, with name agreement of 73% and mostly plausible alternative naming at 19%, and their sets of responses are put through our randomization and analysis protocol, then the mean semantic similarity score is as high as would be expected for the "experimental" randomization, and close to floor for the across-items "baseline control" randomization. This finding acts as a validation of the randomization and semantic analysis protocol. It also provides a helpful context against which to interpret the values to be reported for semantic similarity scores for similar randomizations for meaning-based justifications given to ratings of artworks.

Randomization process for artworks

In the same way as just reported for the calibration analysis, to analyze the responses to the artworks along the same principles, a response from a participant was randomly paired with a response from a different participant so that a score capturing the semantic similarity between the two randomly paired responses could be computed. Again, randomizations were performed using the sequence generator on www.random.org. Randomization outputs were screened for random matches of the responses to themselves, and any of these were re-randomized, so that each response was paired with a different response. As was also the case

for the calibrating analysis, the randomizations for the artworks were just one cycle, not iterative and not bootstrapped. Four different randomizations were performed.

In the first randomization, responses to artworks were randomly paired with other responses to an artwork of the same type (abstract / representational), pooling across all artworks in that class and across all the ages. This formed a baseline control, as it was conceivable that different types of linguistic content would be generated in responses to abstract and representational artworks, which may have led to potentially confounding baseline differences in semantic similarity in random pairs of responses. In the second, the experimental randomization, responses were paired with other responses to the same artwork, again pooling across the age groups. These two randomizations were run to check the main effect of art type on semantic similarity, each against its own baseline.

In the third and fourth randomizations, the grain was finer, in that responses were paired with other responses by a child in the same age group. In the third randomization, the responses were pooled across art type, which formed a baseline control, while in the fourth, the responses were paired with a response to the same artwork in age-specific experimental pairings. These latter two randomizations were included specifically to examine effects of age on the development of shared meaning. They also served as a within-study replication of the findings yielded by randomizations 1 and 2.

For randomizations three and four, there had to be enough responses to each artwork in each age group for responses to be able to be randomly paired with a different response. Bar two instances, this was the case. The two instances where this was not the case were deleted from the analysis in all randomizations to ensure they all drew on the same set of responses. It concerned a semantic response by just one participant to abstract image 11 at

age four, and abstract image 15 at age ten. The remainder of the responses, represented in Table 1, a total of 763, went forward for randomization.

Results

Main Effect of Art Type (Representational vs. Abstract), combining all ages

The first two randomizations, both pooling all ages, were aimed at establishing a main effect of art type on semantic similarity scores. Our first random pairings, randomly pairing a justification response with a different one from the same art type, had the purpose of creating a baseline. The baseline mean for abstract artworks was .30 (median = .31, SD = .17, SE = . 01, 95% CI [.27, .32]) while the baseline mean for representational artworks was a slightly lower .25 (median = .25; SD = .18, SE = ,007 95% CI [.24, .27]). In our second randomization, which produced the experimental pairings, the justifications were randomly paired within the same artwork. In this analysis, abstract artworks had a mean similarity score of .28 (median = .29, SD = .19, SE = .01, 95% CI [.25, .31]), while representational artworks had a higher mean similarity score, namely .36 (median = .35, SD = .20, SE = .008 , 95% CI [.34, .37]). Our focal hypothesis-driven comparison was whether the semantic similarity scores within artworks significantly exceeded the random baseline control pairings within art type. To answer this question, baseline semantic similarity scores were compared to experimental similarity scores, separately for abstract and representational artworks. As with the calibrating analysis, the distributions differed significantly from normal, so the comparisons were made using non-parametric Mann-Whitney tests. For abstract artworks,

while the mean was slightly higher in the baseline control than the experimental condition, there was no significant difference in similarity scores between baseline control pairings and experimental pairings, $Z = -.99$, $p = .32$. However, there was a significant difference when the same comparison was made for representational artworks, $Z = -8.70$, $p < .001$, with the experimental pairings giving rise to mean similarity scores that were approximately .1 higher than those at baseline. This provides evidence that responses to representational artworks had a greater shared meaning than baseline random pairings of the same responses, but that this was not the case for abstract artworks, whose meanings appear to be more idiosyncratic, and therefore did not exceed the semantic similarity scores of baseline pairings. The pattern observed in this set of analyses is somewhat different from the pattern in Schepman et al. (2015a) with adult viewers, where the experimental similarity scores for both representational and abstract artworks exceeded baseline, albeit with a smaller effect size for abstract art.

A reservation about this finding may be found in the imbalance in numbers of justifications that were related to meaning across the abstract and representational artworks. It could be that this lent lower statistical power to the abstract artworks, which could be the reason that their semantic similarity scores did not exceed baseline, while those for the representational artworks did. The viability of this explanation was explored using 172 randomly selected responses from the 589 pairings in the representational sample (matching in number the 172 semantically-based responses to abstract artworks). This was still significant with this smaller sample of 172 artworks from each category, $Z = -3.40$, $p < .001$, making it unlikely that this difference was purely based on differences in statistical power between the two art types.

The next analyses were subsidiary analyses on the full sample, and form analogues to analyses carried out by Schepman et al. (2015a). They were conducted to allow for

comparisons to be made between the two studies. First, we wanted to analyze whether semantic similarity scores differed significantly at baseline for abstract vs. representational artworks to gain a fuller understanding of baseline patterns, which had shown no significant difference in Schepman et al. (2015a). A Mann-Whitney test showed that this baseline difference was significant, $Z = -3.18$, $p = .001$. Thus, at baseline, there was a significantly higher level of semantic similarity for abstract artworks than for representational artworks. As a further analogue to a key analysis in Schepman et al. (2015a) we also compared the semantic similarity scores for the two art types for the experimental pairings. It was found that this experimental difference was also significant, $Z = -3.96$, $p < .001$, with similarity scores being higher for representational than abstract artworks in the experimental condition. This last pattern shows a replication of Schepman et al. (2015a). Importantly, the significant difference in baseline scores shows the need to check baselines and to compare experimental scores against baseline scores for the relevant art type if there is a baseline difference as a function of art type.

Effects of Age on Shared Meaning

Our next set of randomizations, randomizations 3 and 4, was of pairings within age groups, with the third pooling across artworks of the same art type (baseline control), and the fourth using within-artwork pairings (experimental), in an analogue to the first two randomizations described in the previous section, but separated by age group. This set of age-matched pairings was primarily run to examine the effect of age on shared meaning, but also acted as a replication of the main effects reported for randomizations 1 and 2. Statistics for

this analysis are in Table 2. Because the data were complex, and the pattern of means is easier to absorb via a line graph, this is also included, in Figure 1, for the benefit of the reader.

--- INSERT TABLE 2 ABOUT HERE ---

---INSERT FIGURE 1 ABOUT HERE ---

The main reason for running the randomizations with the ages separated was to examine developmental trends on semantic similarity scores, which we hypothesized may be an explanation for the onset of shared liking at age eight observed in Rodway et al. (2016).

We compared, for each age group, separately for abstract and representational art, whether the semantic similarity scores produced by the experimental (within artwork) randomization significantly exceeded the baseline control (within art type, pooled for artworks), using a series of Mann-Whitney tests. The detailed results are presented in the final two columns of Table 2. The pattern was that for each age group as well as for all age groups combined as a main effect, the experimental randomization, when children were talking about the same artwork, exceeded the semantic similarity score observed for the baseline control randomization, but only if the artworks were representational. For abstract artworks, the semantic similarity scores for the experimental randomizations did not exceed the baseline in any age group, nor as a main effect pooling across all groups. First, we note that this set of randomizations, which consisted of entirely new pairings in comparison to those reported in the previous subsection, replicates the main effect reported there, and, again, replicates Schepman et al. (2015a). However, the effect of age showed a finding that is not what was predicted by our interpretation of the Rodway et al. (2016) data, on the basis of

which we might have expected the semantic similarity scores to start exceeding the baseline scores at ages eight and ten, for representation artworks only. The actually observed pattern showed that the shared meaning given to representational artworks was evident from the age of four. We will discuss this main finding more fully in the Discussion section.

Exploring the Baseline Effect: Subsidiary Analyses

It appeared from the baseline measures of central tendency that there may have been baseline effects of age on the semantic similarity score, as the baseline means showed an increase with age. To help us interpret this pattern, by way of subsidiary analysis, we subjected the control randomization data to a Kruskal-Wallis test for differences, which showed a significant effect of age on the baseline semantic similarity scores, $X^2 = 29.27$, df = 3, $p < .001$. We decided to explore potential reasons for this baseline effect, focusing on response length and depth of meaning.

We re-inspected the verbal responses and formed the impression that these were longer in older children. To examine this more robustly, we generated word counts for each individual response included in the current analysis. We found that these data were not normally distributed and therefore analyzed them for age using the non-parametric Kruskal-Wallis test. Pooled across the two art types, the mean length of response rose from 9.40 words (SD = 6.52, SE = .46, 95% CI [8.50, 10.31]) at age four, to 12.57 (SD = 11.52, SE = .84, 95% CI [11.91, 15.24]) at age six, to 23.50 (SD = 15.80, SE = 1.11, 95% CI [21.30, 25.69) at age eight to 25.11 (SD = 12.98, SE = 1.14, 95% CI [22.86, 27.37]) at age ten, significant as a main effect of age $X^2 = 230.47$, df = 3, $p < .001$, with significant increments as pairwise contrasts between ages four and six ($Z = -3.20$, $p < .001$) and six and eight ($Z = -8.50$, $p <$.

001), but not between ages eight and ten ($Z = -1.32$, $p = .19$). The increase in response length with age shows a common fate with the baseline randomization pattern in the semantic analysis data, which showed a similar rise with age. This suggests, on the surface, that the semantic analysis data may be sensitive to the length of the response. The algorithms that are used to generate the semantic similarity scores include sentence-levels elements alongside elements that identify the semantic relatedness of individual words, so it stands to reason that longer sentences could generate higher semantic similarity scores, because they may be more likely to share functional sentence elements.

To examine this further, we correlated the semantic similarity scores from randomization 3 (baseline control, age-matched) against the mean word count of the same two paired responses. This showed a (non-parametric) Spearman correlation coefficient of .30, $N = 761$, $p < .001$. To ensure that we were not simply detecting a correlation attributable to age, which, unsurprisingly, also correlated with similarity scores $\rho = .20$, $N = 761$, $p < .001$ and with word count, $\rho = .65$, $N = 761$, $p < .001$, we also ran a partial correlation between similarity scores and word count, controlling for age, which yielded a coefficient of .21, $df = 758$, $p < .001$. These correlations showed that there was a link between word count and semantic similarity score, but also that this accounted for a relatively small proportion of the overall variance, whether controlling for age (variance accounted for 8.7%) or not (4.6%), suggesting that there must be factors other than basic response length that determine the increase in similarity scores with age.

A further analysis was done to allow for closer interpretation of our pattern of results, this time focusing on depth of meaning. Our data included all responses in which children had been coded as using at least one semantically-based category out of the four that we included in the set (*Subject Matter, Associations, Mood / Emotion, History / Culture*). In the

present analysis, we explored whether the number of semantic categories used by the children (which could range between 1 and 4 in this data set) increased with age and / or varied as a function of art type. We felt this measure may indicate the depth of the semantic interpretation of the artwork. For abstract art, the mean number of semantic comments was 1.09 at age four, 1.19 at age six, and dropped to 1.10 at age eight and 1.00 at age ten, X^2 = 6.19, df = 3, p = .103, not significant as a main effect on a Kruskal-Wallis test, while for representational art we observed 1.17 semantically based comments on average at age four, 1.20 at age six, with a pronounced increase at age eight to 1.32, and 1.29 at age ten, X^2 = 9.86, df = 3, p = .02. When testing for differences in adjacent age groups for the representational artworks only (which was significant as a main effect), the only significant contrast was between ages six and eight, Z = -2.12, p = .03, with the other two contrasts not being significant. We will evaluate in the Discussion whether this step up in the number of semantic categories for representational artworks only at age eight may be related to the onset of shared liking at age eight reported in Rodway et al. (2016).

The Effect of the Presence of Semantic Codes on Aesthetic Appreciation

Our final analysis examined the effect of the presence vs. absence of justifications featuring semantic content on the appreciation of the artworks. It is clear from our data that, despite being non-representational, abstract artworks did attract semantically-based comments. Conversely, representational artworks did not always attract comments which referred to the semantic content of the artwork; instead, some comments reflected only formal artistic properties, color, or other non-semantic aspects of the work. The question we explored

in our final analyses was whether artworks for which there was semantic content in the justifications attracted higher ratings than artworks for which the children did not refer to the semantic content of the artworks. We pitted this analysis against the more traditional division into abstract vs. representational artworks (where the assumption is that semantic content is higher for representational artworks) to explore whether the image-based classification of abstract vs. representational gave the same result as the classification into semantic vs. non-semantic comments based on the children's actual responses. We drew on the entire set of 1600 responses from Rodway et al. (2016) for this set of analyses. Please note that the overall means for abstract vs. representational ratings were previously reported in Rodway et al. (2016), with a minor rounding difference, but analyzed slightly differently there, aggregated by participant, and are therefore reported here again in a comparable disaggregated form to put it on the same footing as all other relevant analyses in this article, to allow for comparisons on the same basis.

As described in more detail above, ratings were given on a scale of 1-5, with 5 being the most favorable. Pooling across abstract and representational artworks in the first instance, for artworks that attracted at least one semantically-based comment, the mean rating was 3.56 (median = 4.00, SD = 1.37, SE = .05, 95% CI [3.46, 3.65]), while for artworks which did not attract any semantically-based comments, the mean was slightly lower at 3.35 (median = 4.00, SD = 1.44, SE = .05, 95% CI [3.25, 3.44]), with these two conditions differing significantly from each other on a non-parametric Mann-Whitney test (used due to ordinal data and non-normality of the distribution), $Z = -2.81$, $p = .005$. Classifying the same data as abstract vs. representational gave rise to a mean of 3.44 (median = 4.00, SD = 1.40, SE = .05, 95% CI [3.34, 3.54]) for representational artworks, and a mean of 3.45 (median = 4.00, SD = 1.41, SE = .05, 95% CI [3.35, 3.55])) for abstract artworks, with the difference between these

two conditions not being significant on a Mann-Whitney test, $Z = -.25$, $p = .80$. In contrast to the notion that representational artworks inherently have content and meaning, this analysis suggests that it is important that the content is noted by the viewer, rather than "inherent" semantic meaning in representational art always automatically leading to higher ratings.

To examine the impact of art type on this observation, the analysis using a division in subjective semantic content was run again, but this time separately for abstract and representational artworks. This revealed a mean rating of 3.35 (median = 4.00, SD = 1.42, SE = .06, 95% CI [3.24, 3.46]) for abstract artworks without subjective semantic content, and a mean rating of 3.81 (median = 4.00, SD = 1.32, SE = .10, 95% CI [3.61, 4.01]) for abstract artworks with semantic comments, a difference which was significant on a Mann-Whitney test, $Z = -3.83$, $p < .001$. For representational artworks, the absence of semantic justifications led to a mean rating of 3.33 (median = 3.00; SD = 1.48, SE = .10, 95% CI [3.13, 3.53]), while their presence led to a slightly higher mean of 3.48 (median = 4.00, SD = 1.38, SE = .06, 95% CI [3.37, 3.59]), but in this case the difference was not significant, $Z = -1.11$, $p = .27$. Thus, interestingly, seeing meaning in the artwork only significantly boosted the ratings of abstract artworks, but not representational artworks.

Discussion

The results allow us to make some observations regarding shared meaning in response to artworks. Considering the main effect of art type on the baseline control vs. experimental semantic similarity scores pooling across the ages (randomizations 1 and 2), based on the patterns of significant differences, there was evidence of shared meaning for representational

artworks, and there was evidence that meaning was not shared significantly above baseline for abstract artworks, with the latter meanings being more idiosyncratic. This confirmed our first hypothesis. The overall pattern of greater shared meaning for representational than abstract art replicates Schepman, et al.'s (2015a) finding using a different sample of participants (children, as opposed to adults in Schepman et al, 2015a), and using different types of verbal materials (longer justification responses in the current data vs. simple associative statements, often consisting of one or a few words in Schepman et al., 2015a). Given the differences in the samples and types of responses across the two studies it is a sign of robustness of the method that the pattern was replicated. The fact that this main effect was replicated again when new randomizations were created to examine the effect of age, boost confidence in the validity further, as does the observation that the effect persisted when the sample sizes for abstract and representational artworks were equated. All these data taken together suggest that the computational semantic analysis method is robust and the data replicable in terms of statistical significance patterns under a varied set of circumstances, strengthening the evidence base for the robustness of this analysis technique.

The actual mean semantic similarity scores need some consideration, too. The numerical differences between experimental and baseline control means were relatively modest in relation to the overall 0-1 scale, despite the robust significance patterns. In the calibration analysis based on Cycowicz et al.'s (1997) picture naming responses the experimental randomization led to a semantic similarity score of .87 and the baseline control . 08, which is a much more pronounced difference between the experimental and baseline control conditions than the figures we observed in the artwork justifications. Recall the baseline control vs. experimental randomizations for abstract artworks was .30 vs.28, respectively, and for representational artworks the means were .25 vs. .36 for baseline control

vs. experimental randomizations, respectively. While these differences are subtle numerically, it is understandable that free-form varied justifications in which children describe why they like an artwork will have less semantic overlap than children naming recognizable objects in single-word responses. It is worth noting that in Schepman et al. (2015a), the mean semantic similarity ratings for associations (single words or short phrases) given to abstract artworks was .07 vs. .11 for baseline vs. experimental randomizations, respectively, and for representational artwork it was .07 vs. .13 for baseline vs. experimental representations, respectively, with differences between baseline and experimental randomizations being significant for both types of art, with a larger effect size for representational art. These differences were even more subtle numerically, and yet also statistically robust. Future testing will need to establish the ranges of numerical values obtained in different tasks and linguistic contexts, before a finer-grained interpretation of the numerical values is fully possible.

In relation to age, our hypothesis was that semantic similarity in the experimental condition would exceed that in the baseline control condition for children aged eight and ten for representational artworks only, because this is where we saw the onset of shared liking for representational art exceeding a shared liking for abstract art (Rodway et al., 2016), and we interpreted (in line with proposals by e.g. Vessel & Rubin, 2010, and Schepman et al., 2015b) that this was due to greater shared meaning having developed by that age. Instead, we discovered that, at all ages from age four upwards, the semantic similarity of responses to representational artworks exceeded baseline. Thus, there was evidence of shared meaning of representational art at younger ages than we had expected based on the interpretation that shared liking was caused by shared meanings, which would take time to develop. The semantic similarity of the verbal responses to abstract art never exceeded baseline, confirming that these varied across individuals at all ages. This is different from the adult

data reported in Schepman et al., (2015a), where the experimental randomizations showed greater semantic similarity scores than baseline controls for abstract artwork, though to a lesser extent than for representational artworks. It is not clear why this discrepancy occurred, and our current data do not permit us to make valid inferences on the potential causes. However, age, response type and response length seem obvious factors to explore in future research.

The finding that shared meaning for representational artworks was present at ages four and six shows that the mere presence of meaning cannot be an interpretation of the shared liking observed to onset from age eight, and that this interpretation needs adjustment. It is possible that the children have shared meaning at younger ages (i.e. ages four and six) than the age at which Rodway et al. (2016) demonstrated shared liking, but that they have more diverse evaluations of these shared meanings at the younger ages, whereas at the older ages, having accumulated more experiences, the emotional and associative aspects of their meanings converge more with those of other children, creating greater shared liking (see Faerber, Leder, Gerger & Carbon, 2010, for a potential mechanism). However, this is a *post hoc* interpretation that would need to be tested more precisely and directly in future research. Alternatively, it is possible that the shared meaning is relatively superficial in young children and it becomes deeper in older children, with superficial shared meaning not leading to shared liking. Such an interpretation would be compatible with our finding that children use significantly more semantic codes in their responses from age eight upwards, which may indicate a greater depth of semantic processing, which, in turn, may be associated with greater shared liking at that age. While such an explanation would have to remain tentative until it were studied more directly in future research, it does chime with proposals by Martindale (1984) and Leder, Belke, Oeberst, and Augustin (2004) that greater meaning

enhances aesthetic appreciation, and thus seems plausible and worthy of further detailed study.

Importantly, as also observed in Rodway et al. (2016), there did not seem to be greater aesthetic appreciation for representational over abstract art in the child sample, which, on the face of it, could be interpreted as meaning not driving aesthetic preferences in children. This could contradict predictions derived from e.g. Landau et al., (2006) and Russell (2003), that representational artworks, which have more inherent meaning, would be preferred over abstract artworks. However, a careful dissection of what constitutes "meaning" enabled us to gain a deeper understanding of this. We discovered that, if children, individually and subjectively, attributed meaning to an artwork, they gave it a higher rating than if they did not use meanings in their justifications for liking. Thus, it would seem that subjective, constructed meaning is of relevance, rather than meaning which is potentially available via representational content in the image. More detailed analysis showed that, while both representational and abstract artworks displayed this phenomenon numerically in their mean ratings, it was only in abstract art that the difference in ratings was significant. This makes sense if one takes the view that the effort to find meaning in an artwork (Russell, 2003) in part determines the hedonic value of finding that meaning, and it may take more effort to identify meaning in abstract than representational artwork, giving rise to a greater sense of reward (Belke, Leder, & Carbon, 2015; Haertel & Carbon, 2014). Overall, the finding that subjective, rather than objective meaning drives art evaluations is important, as it demonstrates that meaning is in the eye of the beholder, and not necessarily automatically present in the image, and that, when meaning is noted, particularly where it is not easy to find, it may make the image more enjoyable to view.

With respect to individual differences in subjective meaning, it would be interesting to further investigate how children's divergent thinking and creative abilities (particularly in narrative storytelling, see Fehr & Russ, 2016, for an example) influence their ratings of both abstract and representational artworks. Children who demonstrate higher levels of divergent thinking and creativity may be more likely to find subjective meaning in both kinds of artwork, but more specifically in abstract artwork through 'romancing' or invention of perceived recognizable content (Winner, 1981) or through the creation of narratives to explain the artwork in an attempt to impose meaning. Personality traits such as extroversion, neuroticism and openness to experience that are related to art preferences (e.g., Furnham & Walker, 2001; Lyssenko, Redies & Hayn-Leichsenring, 2016) and also to levels of creativity (e.g., Burch, Pavelis, Hemsley & Corr, 2006) may also influence the propensity to find subjective meaning, although these relationships have received comparatively little attention in childhood and youth samples.

While meaning may be more idiosyncratic in response to abstract artworks, there are clearly also individual differences in the meaning attributed to representational art. In a key comparison, the representational artworks' similarity scores when paired within images only exceeded the random pairings by .1 (on a range of 0-1, thus 10%). This relatively modest level clearly demonstrates that the overlap observed in the responses to the representational artworks is by no means trivial or obvious, as it could be justifiable to think *a priori*. An inspection of some response pairs illustrates this quite clearly. For example, image 5 depicts a red car on a sunny beach. Two children selected different elements for inclusion in their justification for liking, namely "Because it's like the beach and it's a nice site" and "Because... I like the car on it, it's like Chitty Chitty Bang Bang on the tele", yielding a similarity score of 0.2814. Similarly, two responses to image 6, a dog swimming with a stick

in its mouth gave a similarity score of 0. The responses were "Because it's a doggie and a stick." Paired with "Because he can swim", illustrating that, even if there is content, the justifications do not simply list the key content, but a range of responses is produced, leading to the mean similarity score observed, which is not close to ceiling. This shows that the sharing of meaning is a matter of degree, not a categorical distinction driven by the presence of representational content.

While it was not related to any of our focal hypotheses, we observed that children's semantic similarity scores increased with age as a baseline effect, and we saw a similar increase in the word counts of their responses, consistent with increasing vocabulary acquisition and fluency during the pre-school to primary school years (e.g., Carey, 1978). Further correlation analyses showed that the word count and semantic similarity scores were significantly related, but yet accounted for only a modest amount of shared variance. As for the modest correlation, it may be that the semantic similarity software produced higher scores for longer responses, in part because with more words there is a higher likelihood that the individual words will match, and in part because syntactic structure similarities also feed into the software's algorithms. In itself, this correlation is not of focal interest to the current paper, but it has potentially made our data pattern less clear than it might have been had we known about this in advance. Therefore, in future studies, it may be useful to control the response length, so that semantic similarity can be compared on a similar footing, and particularly so that age groups can be directly compared to each other on an equal basis, something which was not possible in the current data set.

The scope of the current research is also naturally limited by the nature of the sample, which was confined to primary school children. While Schepman et al. (2015a) have explored the phenomenon in adults, further research is warranted to investigate shared meanings in

aesthetic evaluations beyond the current primary school sample, particularly in adolescence where further cognitive maturation towards formal operational functioning (Piaget, 1947) and the influence of more specialist art education (see Burkitt, Jolley & Rose, 2010, Jolley, 2010) may exert an influence. In addition, it could be explored whether social processes or developing personalities may interact in interesting ways with shared aesthetic evaluations during the secondary school years.

Our work has added substantially to our understanding of shared meaning in response to artwork by children aged four to ten. The importance of content, meaning and subject matter had been highlighted by prior researchers (e.g. Leder et al., 2004; Martindale, 1984, Russell, 2003). Content analyses of verbal responses describing the aesthetic experience had also been carried out (Augustin, Carbon & Wagemans, 2012, Augustin, Wagemans & Carbon, 2012), and had been mapped onto visual aspects of artworks, such as color saturation and complexity (Lyssenko et al., 2016). What our work adds is a deeper understanding of the meanings themselves, and the extent to which these are shared across different young viewers.

Conclusions

Our computational analyses have demonstrated that shared meaning is greater in response to representational than abstract art, with semantic similarity scores for the latter not exceeding baseline. Moreover, we have demonstrated that shared meaning of representational artwork is evident in children from age four, which is earlier than the onset of shared liking at age eight (Rodway et al., 2016). Finally, we have shown that the presence of subjective meaning (i.e. meaning found by the viewer, rather than meaning that may be considered to be

inherently present in the image) leads to increased aesthetic ratings, particularly in abstract artworks. Building on Martindale (1984) and Russell (2003), we argue that this is because meaning plays a key role in hedonic value, with additional effort to find meaning potentially giving rise to an enhanced appreciation. The computational semantic similarity analysis used to reach these conclusions has great potential for future research examining the role of meaning in aesthetic appreciation.

References

Augustin, M. D., Carbon, C. C., & Wagemans, J. (2012). Artful terms: A study on aesthetic word usage for visual art versus film and music. *i-Perception*, *3*(5), 319-337.

Augustin, M., Wagemans, J., Carbon, C. (2012). All is beautiful? Generality vs. specificity of word usage in visual aesthetics. *Acta Psychologica, 139,* 187-201.

Augustin, M., & Leder, H. (2006). Art expertise: A study of concepts and conceptual spaces. *Psychology Science, 48,* 135-157.

Augustin, M. D., Leder, H., Hutzler, F., & Carbon, C. C. (2008). Style follows content: On the microgenesis of art perception. *Acta Psychologica*, *128*(1), 127-138.

Belke, B., Leder, H., & Carbon, C. C. (2015). When challenging art gets liked: Evidences for a dual preference formation process for fluent and non-fluent portraits. *PloS one*, *10*(8), e0131796.

Blank, P., Massey, C., Gardner, H., & Winner, E. (1984). Perceiving what paintings express. In W. R. Crozier & A. J. Chapman (Eds.) *Cognitive Processes in the Perception of Art*. 127-143. Amsterdam: Elsevier Science.

Blaye, A., Bernard-Peyron, V., & Bonthoux, F. (2000). Beyond categorisation behaviours: The development of categorical representations between five and nine years of age. *Archives de Psychologie*, 68 (264-265), 59-82.

Burch, S, A., Pavelis, C., Hemsley, D. & Corr, P. J. (2006). Schizotypy and creativity in the visual arts. *British Journal of Psychology*, 97, 177-190.

Burkitt, E., Jolley, J. P., & Rose, S. E. (2010). The attitudes and practices that shape children's drawing experiences at home and at school. *International Journal of Art and Design Education*, 29, 257-270.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller. (Eds.), *Linguistic theory and psychological reality*. (pp. 264-293). Cambridge, MA: MIT Press.

Carothers, T., & Gardner, H. (1979). When children's drawings become art, the emergence of aesthetic production and perception. *Developmental Psychology*, 15, 569-579.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. Psychological Review, 112, 347-382.

Cycowicz, Y. M., Friedman, D., Rothstein, M., & Snodgrass, J. G. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of experimental child psychology*, *65*(2), 171-237.

Eysenck, H. J. (1940). The 'general factor' in aesthetic judgements. *British Journal of*

Psychology, 31, 94-102.

Faerber, S. J., Leder, H., Gerger, G., & Carbon, C. C. (2010). Priming semantic concepts affects the dynamics of aesthetic appreciation. *Acta Psychologica, 135*(2), 191-200.

Fehr, K. K. & Russ, S. W. (2016, June 13). Pretend play and creativity in pre-school age children: Associations and brief intervention. *Psychology of Aesthetics, Creativity and the Arts*, advance online publication.

Freeman, N. H. & Parsons, M. J. (2001). Children's intuitive understanding of pictures. In B. Torff & R. J. Sternberg (Eds.), *Understanding and teaching the intuitive mind.* (pp.73-91). Mahwah, NJ: Erlbaum.

Gardiner, H., Winner, E., & Kirchner, M. (1975). Children's conceptions of the Arts. *Journal of Aesthetic Education*, 9, 60-77.

Furnham, A., & Walker, J. (2001). Personality and judgements of abstract, pop art, and representational paintings. *European Journal of Personality*, 15. 57-72.

Haertel, M., & Carbon, C. C. (2014). Is This a "Fettecke" or Just a "Greasy Corner"? About the Capability of Laypersons to Differentiate between Art and Non-Art via Object's Originality. *i-Perception*, *5*(7), 602-610.

Han L., Kashyap A., Finin T., Mayfield J., Weese J. (2013). *UMBC EBIQUITY-CORE: Semantic textual similarity systems*. Retrieved from http://ebiquity.umbc.edu/paper/html/id/621

Jolley, R. P. (2010). *Children and pictures: drawing and understanding*. Oxford: Wiley-Blackwell.

Koch, A., Alves, H., Krüger, T., & Unkelbach, C. (2016). A general valence asymmetry in similarity: Good is more alike than bad. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1171 – 1192.

Landau, M. J. , Greenberg, J., Solomon, S., Pyszczynski, T. & Martens, A. (2006). Windows into nothingness: Terror management, meaninglessness, and negative reactions to modern art. *Journal of Personality and Social Psychology*, 90 (6), 879-892.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

Leder, H., Belke, B., Oeberst, A., and Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95, 489–508.

Leder, H., Gerger, G., Dresser, S. G. & Schabmann, A. (2012). How art is appreciated. *Psychology of Aesthetics, Creativity, and the Arts*. Vol 6(1), 2-10. doi: 10.1037/a0026396

Leder, H., Goller, J., Rigotti, T., & Forster, M. (2016). Private and Shared Taste in Art and Face Appreciation. *Frontiers in human neuroscience*, *10*.

Lin, S.F., & Thomas, G.V. (2002). Development of understanding of popular graphic art: a study of everyday aesthetics in children, adolescents and young adults. *International Journal of Behavioral Development,* 26, 278–287.doi:10.1080/016502501430 00157

Locher, P., Krupinski, E. A., Mello-Thoms, C., & Nodine, C. F. (2007). Visual interest in pictorial art during an aesthetic experience. *Spatial Vision*, *21*(1), 55-77.

Lyssenko, N., Redies, C. & Hayn-Leichsenring, G. U. (2016). Evaluating abstract art: Relation between term usage, subjective ratings, image properties and personality traits. *Frontiers in Psychology*, 7 (973) 1-9.

Machotka, P. (1966). Aesthetic criteria in childhood: justifications of preference. *Child Development.* 37, 877–885. doi:10.2307/1126610

Martindale, C. (1984). The pleasures of thought: A theory of cognitive hedonics. *Journal of Mind and Behavior*, 5(1), 49-80.

McManus I.C. (1980). The aesthetics of simple figures. *British Journal of Psychology*, 71, 502–24.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):41.

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA, MIT Press.

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual review of psychology*, *64*, 77-107.

Parsons, M. J. (1987). *How we Understand Art: A Cognitive Developmental Account of Aesthetic Experience.* Cambridge: Cambridge University Press.

Pelowski, M., Markey, P. S., Lauring, J. O., & Leder, H. (2016). Visualizing the impact of art: An update and comparison of current psychological models of art experience. *Frontiers in human neuroscience*, *10*. http://journal.frontiersin.org/article/10.3389/fnhum.2016.00160

Piaget, J. (1947). *The psychology of intelligence*. London: Routledge and Kegan Paul Ltd.

Potter, T., Corneille, O., Ruys, K., & Rhodes, G. (2007). S/he's just another pretty face: A multidimensional scaling approach to face attractiveness and variability. *Psychonomic Bulletin & Review, 14*, 368–372.

Qin, S., Cho, S., Chen, T., Rosenberg-Lee, M., Geary, D. C., & Menon, V. (2014). Hippocampal-neocortical functional reorganisation underlies children's cognitive development. *Nature Neuroscience*, 17, 1263-1269.

Rodway, P., Kirkham J., Schepman, A., Lambert, J., Locke, A. (2016). The development of shared liking of representational but not abstract art in primary school children and their justifications for liking. *Frontiers in Human Neuroscience*, 10 (21).

Russell, P. A. (2003). Effort after meaning and the hedonic value of paintings. *British Journal of Psychology*, *94*(1), 99-110.

Schabmann, A., Gerger, G., Schmidt, B.M., Wögerer, E., Osipov, I. & Leder, H. (2016).Where does it come from? Developmental aspects of art appreciation. International Journal of Behavioral Development. 40(4) 313-323.

Schepman, A. Rodway, P. & Pullen, S. J. (2015a). Greater cross-viewer similarity of semantic associations for representational than for abstract artworks. *Journal of Vision*, 15(14):12, 1–6.

Schepman, A., Rodway, P., Pullen, S. & Kirkham, J.A. (2015b). Shared liking and association valence for representational art but not abstract art. *Journal of Vision*, 15(5), 11: 1-10.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, *6*(2), 174 - 215.

Tinio, P. P. L. (2013). From artistic creation to aesthetic reception: The mirror model of art. *Psychology of Aesthetics, Creativity and the Arts*, 7, 265-275.

Toomela, A. (2002). Drawing as a verbally mediated activity: A study of relationships between verbal, motor and visuospatial skills and drawing in children. *International Journal of Behavioural Development*, 26 (3), 234-247.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M., & Danner, D. (2008). Why positive information is processed faster: the density hypothesis. *Journal of personality and social psychology*, *95*(1), 36 – 49.

Vessel, E. A., & Rubin, N. (2010). Beauty and the beholder: highly individual taste for abstract, but not real-world images. *Journal of Vision*, *10*(2), 18-18.

Vo N. and Popescu O. (2016). A Multi-Layer System for Semantic Textual Similarity. In Fred, A., Dietz, J., Aveiro, D., Liu, K., Bernardino, J. & Filipe, J.(Eds.), *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016).* Setúbal, Portugal: Scitepress, Science and Technology Publications, Lda ISBN 978-989-758-203-5, 56-67. DOI: 10.5220/0006045800560067

Winner, E. (1982). *Invented worlds: The psychology of the arts.* Cambridge MA: Harvard University Press.

FIGURE 1:

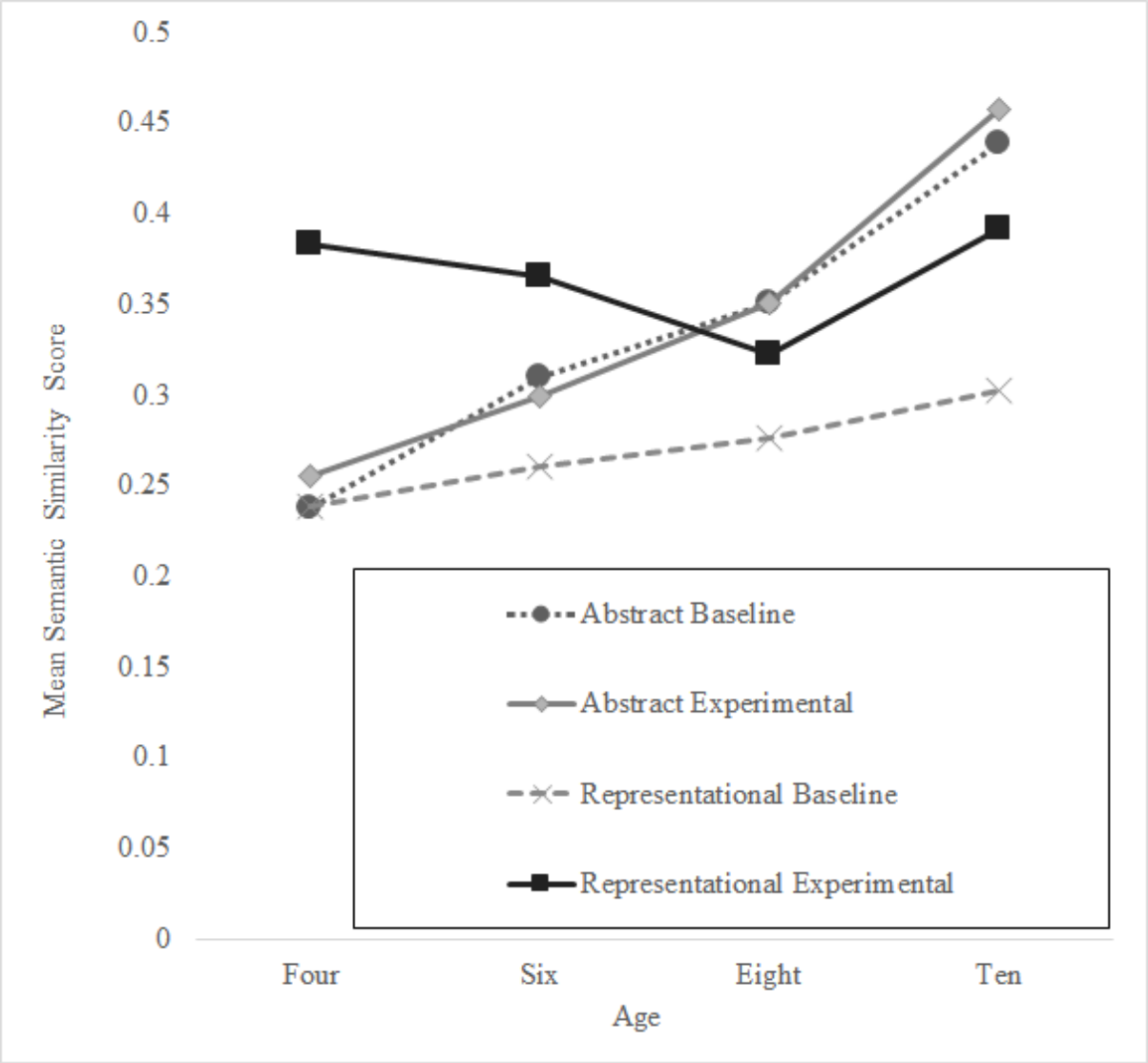

Figure 1: Line graph of mean semantic similarity scores for randomizations 3 (baseline control, across artworks) and 4 (experimental, within-artwork), separated by age and art type.

Table 1

Numbers and percentages of cases with semantically-based justifications

Age	Abstract		Representational		Total	
	Count	%	Count	%	Count	%
Four	47	24	156	78	203	51
Six	43	22	143	72	186	47
Eight	55	28	147	74	202	51
Ten	29	15	143	72	172	43
Total	174	22	589	74	763	48

Note: The non-semantically-based comments for abstract and representational are the complement to 200, for total per age group to 400, total per art type to 800, with 1600 data points overall. In columns marked %, the number of observations is expressed as a percentage of the number of opportunities. Note that, in the final analysis, N was reduced by 1 for Abstract artworks at ages four and ten, due to single responses to an artwork not being able to be paired with a different response to the same artwork.

Table 2

Means, Medians, SDs, SEMs and 95% Confidence Intervals for semantic similarity scores for randomizations 3 (baseline randomization within art type, separated by age) and 4 (experimental randomization within artwork, separated by age).

		Baseline Control					Experimental					Diffrence	
Age	Art type	Mean	Median	SD	SEM	95% CI	Mean	Median	SD	SEM	95% CI	Z	p
Four	A	.24	.14	.25	.04	.16, .31	.26	.16	.24	.04	.18, .33	-0.53	0.60
	R	.24	.23	.21	.02	.20, .27	.38	.36	.24	.02	.35, .42	-5.41	< .001
	B	.24	.22	.22	.02	.21, .27	.35	.33	.24	.02	.32, .39		
Six	A	.31	.31	.23	.03	.24, .38	.30	.29	.21	.03	.24, .36	0.00	1.00
	R	.26	.26	.19	.02	.23, .29	.37	.36	.21	.02	.33, .40	-4.29	< .001
	B	.27	.28	.20	.01	.24, .30	.35	.35	.21	.02	.32, .38		
Eigh t	A	.35	.34	.17	.02	.30, .40	.35	.40	.18	.02	.30, .40	-0.32	0.75
	R	.28	.30	.15	.01	.25, .30	.32	.34	.16	.01	.30, .35	-2.36	0.02
	B	.30	.31	.16	.01	.27, .32	.33	.35	.17	.01	.31, .35		
Ten	A	.44	.44	.14	.03	.38, .50	.46	.45	.17	.03	.40, .52	-0.43	0.67
	R	.30	.30	.16	.01	.28, .33	.39	.41	.18	.02	.36, .42	-4.37	< .001
	B	.32	.32	.16	.01	.30, .35	.40	.42	.18	.01	.38, .43		
All	A	.33	.34	.22	.02	.29, .37	.33	.35	.21	.02	.30, .36	-0.53	0.60
	R	.27	.28	.18	.01	.25, .28	.37	.36	.20	.01	.35, .38	-8.25	< .001
	B	.28	.29	.19	.01	.27, .29	.36	.36	.20	.01	.34, .37		

Note: A = Abstract, R = Representational, B = Both art types. Z and p from Mann-Whitney tests comparing the baseline vs experimental randomization for each art type are in the final two columns.

Appendix: Artworks Used in the Study

The formally published version of this appendix is available via:
http://dx.doi.org/10.1037/aca0000159.supp

Representational Artworks

1: Kevin Heaney: Houses Granite Montana

2: Ian Sheldon: Peeling Wallpaper

3: David Wade: Streamside

4: Bruce Greene: Under the Indian Blanket

5: Mark Peterson: '55 T-Bird

6: Jay Kemp: Return to Sender

7: Sergio Zampieri: Autumn Light

8: Albert Edelfelt: Boys Playing on the Shore

9: Jean Smith: Laugher #4

10: Paul Dixon: Ups and Downs

Abstract Artworks

11: Boi K' Boi: Mah Abstract Colors Niamh

12: Unknown Artist: Ode to Miro

13: Mystral Casterial: Kandinsky Tribute

14: Elizabeth Urabe: In God's Hands

15: Stephanie Kordan Dardashti: Desire Red

16: Mauren Greenwood: Indulgence

17: Brice Marden: Cold Mountain

18: ScentOfBlood: Kandinsky Tribute

19: Ingrid Claessen: Nature Green Yellow White

20: Ingrid Claessen: No4